Fostering Trust through Gesture and Voice-Controlled Robot Trajectories in Industrial Human-Robot Collaboration

Campagna, Giulio; Frommel, Christoph ; Haase, Tobias; Gottardi, Alberto; Villagrossi, Enrico; Chrysostomou, Dimitrios; Rehm, Matthias

Published in:
2025 International Conference on Robotics and Automation (ICRA)

Publication date:
2025

Document Version
Accepted author manuscript, peer reviewed version

Fostering Trust through Gesture and Voice-Controlled Robot Trajectories in Industrial Human-Robot Collaboration*

Giulio Campagna[1], Christoph Frommel[2], Tobias Haase[2], Alberto Gottardi[3], Enrico Villagrossi[4], Dimitrios Chrysostomou[5], Matthias Rehm[1]

***Abstract*— In the Industry 5.0 era, the focus shifts from basic automation to fostering collaboration between humans and robots. Trust is crucial in this new paradigm, enabling smooth interaction, especially for users with limited robotics knowledge. This study presents a novel framework that uses human hand gestures and voice commands to control robot movements, aiming to enhance trust, reduce cognitive workload, and minimize task execution time—key for efficient manufacturing. In automated systems, swift completion of micromanagement tasks is essential to prevent process disruption. To evaluate this framework, we devised a testbed scenario within an automated carbon fiber transportation and draping process, focusing on a maintenance task as the micromanagement challenge. Participants inspected the gripper, guided the robot along a defined path, and performed maintenance, such as attaching cables. Two conditions were tested: gestures and voice commands versus a smartPAD. The results showed that gestures and voice commands increased trust, lowered cognitive load, and shortened execution times, improving overall manufacturing efficiency.**

I. INTRODUCTION

Industry 5.0 era transformed traditional manufacturing by adopting a human-centric approach that integrates advanced technologies. This shift requires a redefinition of human-robot collaboration (HRC), where trust and cognitive workload are critical for safe and effective operations [1]. As collaborative robots become more common, maintaining trust is essential for productive interactions and operator safety [2]–[4].

In this study, we use Muir and Moray's definition of trust [5], which is the operator's confidence in a system's competence and reliability to accomplish the task. In industrial HRC, trust affects interaction quality, especially for operators with limited robotic experience [6]. Additionally, cognitive workload – the mental effort required for task performance – is crucial for the effectiveness and safety of collaborative processes [7], [8]. Previous research has investigated various methods to enhance trust and reduce cognitive load in HRC. Studies have investigated how robot appearance [9], [10], behavior [11], and communication strategies [12] influence human trust. Additionally, research focused on adaptive control systems [13] and real-time psychophysiological monitoring [14] to optimize robot behavior. Despite these advances, many solutions involve complex interfaces or extensive operator training, limiting their industrial applicability. Recent research has delved into more nuanced trust aspects in HRC. For instance, Campagna et al. explored using facial features for trust evaluation [15], while another study examined body motion data as a trust indicator [16]. Although these methods offer real-time trust assessment, they may not directly address intuitive robot control. To address these issues, studies have analysed more natural interaction modalities. Gesture-based control systems have been effective in improving operator comfort and reducing cognitive load [17], while voice command interfaces have shown potential for enhancing communication efficiency in HRC [18]. However, these approaches are often used in isolation, potentially missing the benefits of multimodal interaction.

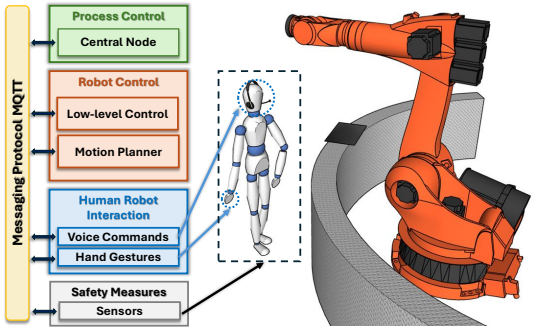

Fig. 1: Overview of the proposed framework to promote trust and improve the manufacturing process in industrial HRC.

In this paper, we present a novel framework that integrates hand gestures and voice commands for controlling robot trajectories, aimed at enhancing trust and reducing cognitive load in HRC. This approach merges these intuitive interaction modalities into a holistic system, providing a more flexible and user-friendly control paradigm. By combining spatial gesture control with precise verbal instructions, our framework seeks to create a more accessible and trust-enhancing interface for collaborative robots. It uses the *Message Queue Telemetry Transport* (MQTT) protocol for efficient data exchange, ensuring smooth integration of gesture recognition, voice command processing, and robot control systems (see Figure 1). To evaluate our approach, we developed a testbed scenario

* The work was supported by the European Union's Horizon 2020 research and innovation program (Grant No. 101006732), and the Independent Research Fund Denmark (Grant No. 1032-00311B).

Corresponding author's email: gica@create.aau.dk

[1] G.Campagna and M.Rehm are with the Human-Robot Interaction Lab., Technical Faculty of IT and Design, Aalborg University, Denmark.

[2] C.Frommel and T.Haase are with the Center for Lightweight Production Technology, German Aerospace Center (DLR), Augsburg, Germany.

[3] A.Gottardi is with the IT+Robotics s.r.l and the Intelligent Autonomous System Lab, Department of Information Engineering, University of Padua, Padua, Italy.

[4] E.Villagrossi is with the Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing, CNR-STIIMA, Milan, Italy.

[5] D.Chrysostomou is with the Smart Production Lab., Faculty of Engineering and Natural Sciences, Aalborg University, Denmark.

centered around a maintenance task—an example of a micromanagement operation—following the transport and draping of carbon fiber fabric. The human operator guides the robot along a predefined trajectory to a target position for maintenance activities, such as attaching a cable to the robot's gripper. This setup underscores the importance of trust and clear communication to avoid potential risks like collisions or misalignment.

The framework was evaluated through a comparative study involving two conditions: *traditional smartPAD control* and the *gesture and voice-assisted system*. Performance metrics included task execution time, trust levels measured by the *Trust Perception Scale-HRI* [19], and cognitive workload assessed using the *NASA TLX questionnaire* [20]. As an additional analysis, we evaluated perceived intelligence and safety using the *Godspeed Questionnaire* [21]. This evaluation provides empirical evidence on how our multimodal interaction approach impacts trust, cognitive workload, and task efficiency in HRC.

The key contributions of this paper are:

- A novel framework integrating gesture and voice control for robot trajectory guidance in industrial HRC.
- A flexible system that allows for seamless micromanagement without disrupting automated processes.
- Empirical evidence on the impact of multimodal interaction on trust, cognitive workload, and task efficiency in HRC.
- Guidelines for designing intuitive and trust-building interfaces for collaborative robots in Industry 5.0.

II. METHODOLOGY

This work presents a framework that integrates gesture recognition, voice commands, and adaptive robot control to create a user-friendly interface for operators with limited robotics experience. It is important to highlight that the framework's novelty is in its facilitation of seamless collaboration, with detailed benefits discussed in Section IV and Section V.

A. Hand Gestures Detection

For reliable **hand tracking**, we employed *MediaPipe Hands* [22], known for its accuracy and ease of integration. This real-time system analyzes RGB images to predict the hand skeleton, handling various hand appearances, sizes, lighting conditions, and backgrounds. It accurately identifies key landmarks such as fingertips, knuckles, and wrist points, providing essential spatial data for gesture recognition. The model pipeline includes: i) a *palm detector* using an oriented bounding box to locate palms, and ii) a *hand landmark* model that provides precise 2.5D landmark coordinates from the cropped bounding box.

Palm Detector: A Single-Shot Detector is employed to identify palms, which are easier to detect than full hands with fingers due to the lack of distinctive features. Trained on 6,000 images, the palm detector uses a *Feature Pyramid Network* and focal loss optimization to handle varying scales and numerous anchors. Feeding the landmark model with precisely cropped hand images minimized the need for data augmentation and enhanced the accuracy of coordinate predictions.

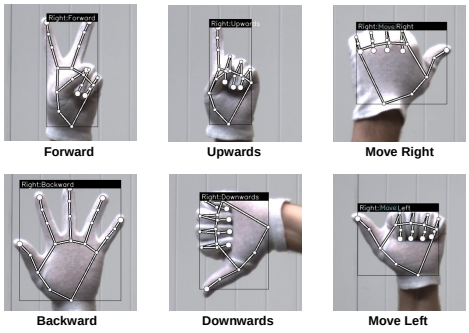

Fig. 2: The six right-handed gestures, viewed from the camera, control the robot's direction. For safety, the human operator is required to wear gloves.

Hand Landmark Model: The model predicts 21 2.5D hand keypoints, including depth relative to the wrist, using a regression-based approach. Leveraging a topology similar to *Multi-view Bootstrapping* [23], it reliably represents hand poses, even under partial obscuration and self-occlusion. The model also outputs hand presence probability and handedness classification. Trained on around 30,000 real-world images annotated with 21 keypoints, it uses synthetic hand models for enhanced pose coverage and geometric supervision. Real-world images provide annotated landmarks, while synthetic images are mapped to ground-truth 3D joints. Hand presence detection uses positive samples from annotated images and negatives from non-annotated areas, with handedness identified by labeling a subset of images as left or right hands.

Following *hand tracking*, a **gesture recognition** model was developed utilizing a *Dense Neural Network* (DNN), inspired by [24]. This model identifies six right-handed gestures for robot control (see Figure 2). Hand landmark data were collected via MediaPipe, which provided 21 (x, y, z) normalized coordinates from images captured at 24 Hz and a resolution of 1368x912. The dataset comprised 9,632 samples, distributed across six gesture classes, and was partitioned into 60% training, 20% testing, and 20% validation sets. Multiple DNN architectures were evaluated, with variations in layers, neurons, dropout rates, optimizers, and learning rates. The most effective *DNN architecture* begins with an *input layer* that processes input vectors representing hand landmark coordinates. To prevent overfitting, a *dropout layer* with a *0.2 dropout rate* is applied. This is followed by a *dense layer* with *32 neurons* and *ReLU* activation, another *dropout layer* with a *0.4* dropout rate, and a *dense layer* with *16 neurons* and *ReLU* activation. The final *output layer* consists of *6 neurons* with *softmax* activation. The model was trained using the *Adaptive Moment Estimation* optimizer, with a *learning rate* of *0.001*, *beta1* of *0.9*, and *beta2* of *0.999*. *Sparse Categorical Cross Entropy* served as the loss function, and an *early stopping criterion* with a *patience* of *20 epochs* was employed. Training stopped after *181 epochs*, using a

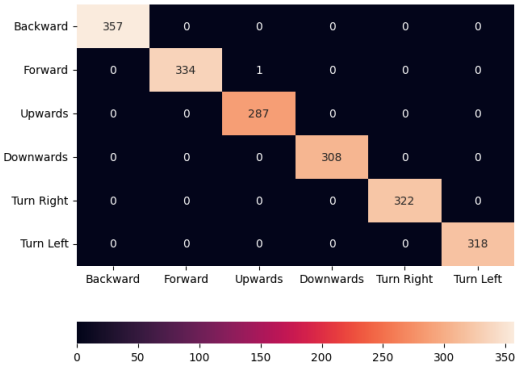

Fig. 3: Confusion matrix for the gesture recognition model, showing a perfect classification accuracy across all six gesture categories with a single misclassification.

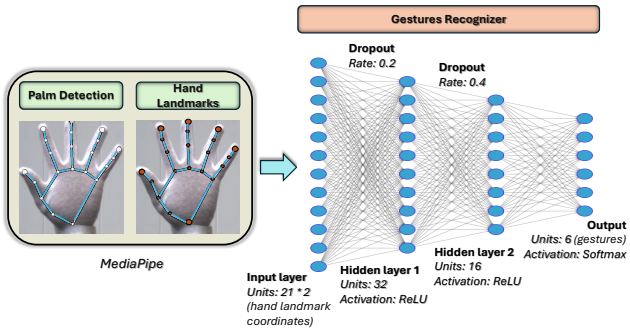

Fig. 4: Overall framework for the hand gesture detection.

batch size of *32*. The model achieved a 99.95% classification accuracy, with precision, recall, and F1-score all reaching 100%. Figure 3 shows the confusion matrix. To improve gesture predictions, class occurrences were counted over a 1-second interval with a detection confidence threshold of 0.7. The class with the highest frequency was selected as the predicted gesture. Figure 4 depicts the overall framework of the hand gesture detection model.

B. Voice Recognition System

The framework in Figure 1 features a *voice assistant* that specifies the distance the robot should cover based on the direction indicated by hand gestures and provides feedback on voice command recognition. We use *Rhasspy*[1], an open-source toolkit for custom voice interfaces. *Rhasspy* provides personalized wake words, speech-to-text, and text-to-speech models, with modular design and multilingual support, making it versatile for various applications. Unlike many commercial voice assistants, *Rhasspy* processes everything locally, ensuring data privacy. In the following, it is detailed the framework's key concepts, components (see Figure 5) and the underlying mechanism.

The process starts when the human operator speaks into the system's microphone. *PyAudio* captures and streams the audio directly. The system listens for the wake word "*Porcupine*," and upon detection, activates and processes the subsequent audio, which is then sent to *Kaldi* for *Automatic Speech Recognition* (ASR) with a minimum confidence threshold of 0.7. Although trained in English, the ASR system can be retrained for other languages. In this study, the operator uses voice commands to control the robot's movement through a Natural Language Understanding (NLU) system that interprets various commands, focusing on syntaxes such as "*Move the robot [distance] [unit of measure]*" or "*Move [distance] [unit of measure]*" to allow flexible phrasing and optional components. *Kaldi* converts spoken words into text, which is processed by Rhasspy's intent recognition engine, *Fsticuffs*. *Fsticuffs* uses NLU to match the transcribed text to predefined commands and employs fuzzy text matching for improved accuracy, handling minor variations in user input. Upon recognizing the intent, the *Dialogue Manager* triggers a custom application to determine the appropriate action, such as starting the robot's motion. This application collects data, fills in missing information with contextual clues, and requests additional input if necessary [25]. The system then provides feedback to the operator, confirming if the command was recognized. If successful, it sends the target pose to the *Motion Planner* and executes the motion. If not, it identifies the issue (e.g., incorrect pronunciation) and prompts the user to retry. Feedback is delivered via synthetic speech generated by *NanoTTS* and audio playback managed by *ALSA* (Advanced Linux Sound Architecture), completing the interaction loop. Communication among components is facilitated by *MQTT* and *Hermes* protocols. *MQTT* ensures lightweight, real-time messaging, while *Hermes* manages dialogue flow, integrating tasks such as intent recognition, command execution, and feedback delivery.

[1]Rhasspy Voice Assistant: https://rhasspy.readthedocs.io/en/latest/

C. Robot Control and Planning

Figure 1 shows the robotic cell's control framework, featuring a *Central Node* [26] that integrates robot control with human command recognition. This node uses *environmental sensors*, including laser scanners, to monitor the work cell, create 3D maps, and detect individuals. If the protective zone is breached, the robot stops to avoid collisions or injuries. The *Central Node* processes the operator's position and commands (e.g., gestures, voice) and relays these data to the *Robot Control* module. This module includes a *Motion Planner* for trajectory generation and *Low-Level Control* that communicates via the real-time *KUKA Robot Sensor Interface* (RSI) to safely execute commands.

The *Motion Planner* algorithm used is a modified version of *Rapidly-exploring Random Tree Connect* (RRT-Connect) [27], tailored for safe collaboration with human operators. It relies on two search trees, T_{start} and T_{goal}, which expand from the starting configuration q_{start} and goal configuration q_{goal}, respectively. The key steps of the algorithm can be summarized in two main equations:

1. Finding the nearest node:

$$q_{near} = \arg\min_{q \in T} \|q - q_{rand}\| \quad (1)$$

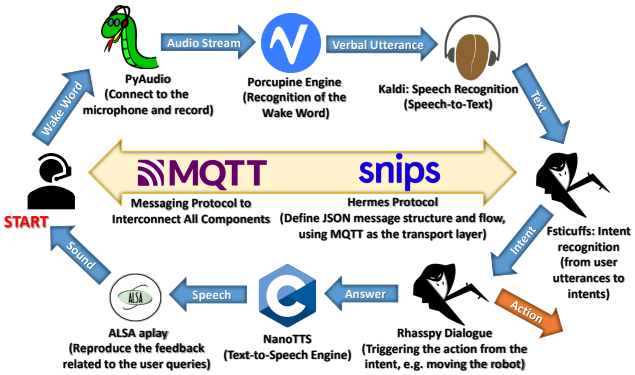

Fig. 5: Components of the Rhasspy voice assistant and the underlying mechanism.

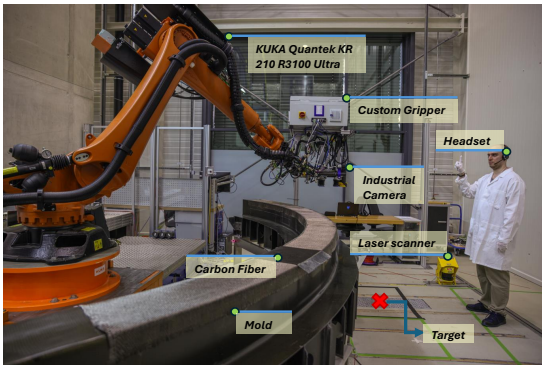

Fig. 6: The experimental scenario. The operator may use hand gestures and voice commands, or a smartPAD, depending on the experimental condition.

where q_{rand} is a randomly sampled configuration, and T is either T_{start} or T_{goal}.

2. Extending the tree:

$$q_{new} = q_{near} + \min(\varepsilon, \|q_{rand} - q_{near}\|) \cdot \frac{q_{rand} - q_{near}}{\|q_{rand} - q_{near}\|} \quad (2)$$

where ε is the maximum step size, ensuring that the new node q_{new} is within a fixed distance from q_{near}.

The algorithm alternates expanding the two trees until they connect, forming a continuous path. RRT-Connect is preferred for its bidirectional growth, which speeds up pathfinding and reduces search time compared to standard RRT. To ensure safety, a collision-free volume around the trajectory is computed using the method described in [28] and then sent to the *Low-Level Control* for execution.

The *Low-Level Control* module uses the *ros_control* approach, a standard in robotics that ensures consistent interfaces for low-level controllers and hardware. This approach enables real-time hardware communication through custom drivers while maintaining standardized higher-level control processes like path planning, simplifying integration and maintenance. The module is managed by two PCs: a Beckhoff Industrial PC (IPC) and a Linux PC running ROS. The Beckhoff IPC, chosen for its real-time interface capabilities, uses TwinCAT software to provide real-time performance and run PLC programs in IEC-61131-3 and C++. This setup consolidates automation software and efficiently controls the robotic cell. A custom RSI driver in TwinCAT manages operations and executes trajectories from the external motion planner. The Linux PC handles the nominal trajectory from the *Motion Planner*, micro-interpolating the trajectory at 4 ms intervals to comply with the RSI control interface. Using two PCs is essential for performance: the Beckhoff IPC handles real-time robot communication, while the Linux PC executes other control algorithms.

III. EXPERIMENTS

This section describes the experimental procedure used to evaluate the proposed framework's effectiveness in enhancing trust, reducing cognitive workload, and improving the manufacturing process in industrial HRC.

A. Task Description and Experimental Conditions

An industrial robot transports carbon fiber fabric to a mold, where a human operator drapes the material with the robot's assistance to ensure it fits the mold. Both processes are automated but must be paused for micromanagement tasks such as maintenance checks, requiring manual robot control via a smartPAD, which disrupts the manufacturing workflow. To address this challenge, we developed a framework for robot control using hand gestures and voice commands. This approach ensures uninterrupted industrial processes and is designed for operators with limited robot control experience. In contrast, traditional smartPADs can complicate control due to their unfamiliarity, highlighting the need for more intuitive interfaces.

To validate this framework, we designed a scenario where the operator performs a *maintenance task*. After inspecting the gripper, the operator moves the robot to a target location (see Figure 6) away from the carbon fiber to prevent damage. We tested two conditions: (*i*) robot control via smartPAD and (*ii*) robot control using hand gestures and voice commands. In both cases, the robot follows the same path: it moves 260 cm right, 100 cm forward, and approximately 30 cm downward (adjusted for operator height). The operator then attaches a cable to the gripper and records the inspection date. The robot is then moved upward by the same distance and 100 cm backward to return to the mold, preparing for the next transport and draping cycle. Hand gestures direct the robot's motion (see Figure 2), while voice commands specify distances (e.g., "*Move the robot 100 centimeters*"). This dual-modality approach enhances intuitive interaction and aligns the operator's spatial perception with the robot's programmed reference frame, ensuring more precise and seamless control in dynamic settings.

B. Experimental Setup

The experimental setup uses a *KUKA Quantek KR 210 R3100 Ultra* robot with a *KR C4 controller*. This robot, featuring six revolute joints, a 3095 mm reach, and an external

linear axis, is equipped with a *custom gripper*[2] that employs suction units based on the Coanda effect for secure gripping. The *Jabra headset* provides high audio quality and Active Noise Cancellation for voice commands, while an *industrial camera* (IDS Imaging Development Systems GmbH) captures hand gestures with a GigE interface, and is configured to 1368x912 pixels, 24 fps, and 40.31 ms exposure time. The camera is mounted on the end-effector at an optimal height and angle for clear imaging.

Safety measures include laser scanners to detect obstacles, emergency stop buttons, and protective gear for participants (safety shoes, lab coats, gloves). Collision-free algorithms (see Section II-C) and safety zones minimize risk, with the robot's speed limited to 250 mm/s per ISO 10218-1:2011 standards. All participants receive thorough training on safety protocols. The setup is illustrated in Figure 6.

C. Experimental Protocol

The experiment was conducted with 22 healthy participants (16 males, 6 females) of varying ages ($M = 35.45$, $SD = 8.62$), all recruited from the personnel at the DLR - German Aerospace Center. Participants were selected specifically for their limited prior experience with robotics.

The experimental protocol adhered to the *Declaration of Helsinki* and received institutional review board ethical approval. Participants were given a consent form outlining the study's objectives, tasks, risks, methods, and potential benefits. They reported their gender and age, then completed the 6-item *Propensity to Trust Questionnaire* [29] to establish baseline trust in the system. Protective equipment was then donned, and participants completed two trials per randomized condition. After each condition, they filled out the 14-item *Trust Perception Scale-HRI* [19] to measure trust in the system, and assessed *Perceived Intelligence* and *Perceived Safety* using the *Godspeed Questionnaire* [21], which were deemed relevant for further investigation in line with the study's objectives. Participants also completed the *NASA TLX Questionnaire* [20] to evaluate cognitive workload. Responses were normalized to a $[0,1]$ scale. Task execution time for participants with limited robotics knowledge was recorded only for the robot control segment of each trial, excluding gripper and cable inspection. The total participant involvement was 90 minutes, including the introductory briefing.

IV. EXPERIMENTAL RESULTS

Participants' *propensity to trust* was evaluated, resulting in a high average score ($M = 0.80$, $SD = 0.06$), with individual scores ranging from 0.71 to 0.92. Therefore, the participant pool was deemed consistently inclined to trust the system, removing the need for distinctions in later analyses. *Trust* in the system was analyzed based on the within-subjects design across two conditions: (*i*) robot control using the smartPAD and (*ii*) robot control via hand gestures and voice commands. The *Shapiro-Wilk test* confirmed a normal distribution of the difference in average trust scores between these modalities ($p = 0.193$). A *paired t-test* revealed significantly higher trust scores for the hand gestures and voice commands condition ($M = 0.90$, $SD = 0.08$) compared to the smartPAD condition ($M = 0.60$, $SD = 0.07$), with ($t(21) = 12.572$, $p = 3.068 \cdot 10^{-11}$). Additionally, we investigated participants'perceptions of the system's intelligence and safety. In examining *perceived intelligence*, the *Shapiro-Wilk test* indicated that the differences in scores between the two control modalities did not follow a normal distribution ($p = 0.046$), prompting the use of the *Wilcoxon Signed-Rank test*. The statistical analysis revealed a significant difference in perceived intelligence scores between the hand gestures and voice assistant scenario ($M = 0.85$, $SD = 0.09$) and the smartPAD scenario ($M = 0.43$, $SD = 0.11$). Specifically, the perceived intelligence score for the hand gestures and voice assistant scenario ($Mdn = 0.85$) was significantly higher than that for the smartPAD scenario ($Mdn = 0.45$), with $p = 2.384 \cdot 10^{-7}$. For *perceived safety*, the *Shapiro-Wilk test* showed that the differences in scores between the two control modalities were normally distributed ($p = 0.390$). A *paired t-test* found significantly higher safety scores for the hand gestures and voice commands scenario ($M = 0.85$, $SD = 0.09$) versus the smartPAD scenario ($M = 0.46$, $SD = 0.11$), with $t(21) = 13.3$, $p = 1.068 \cdot 10^{-11}$. Regarding *cognitive workload*, the *Shapiro-Wilk test* indicated that the differences in workload scores followed a normal distribution ($p = 0.075$). A *paired t-test* indicated that cognitive workload was significantly higher in the smartPAD scenario ($M = 0.46$, $SD = 0.13$) compared to the hand gestures and voice commands scenario ($M = 0.09$, $SD = 0.03$), with $t(21) = 13.206$, $p = 1.221 \cdot 10^{-11}$. Finally, *task execution time* was analyzed, with the *Shapiro-Wilk test* confirming that the differences in execution times followed a normal distribution ($p = 0.181$). A *paired t-test* showed that the average execution time with the smartPAD ($M = 288.45$, $SD = 69.58$) was significantly longer than with hand gestures and voice commands ($M = 124.20$, $SD = 20.63$), with $t(21) = 10.581$, $p = 7.141 \cdot 10^{-10}$. Figures 7 and 8 present the summarized questionnaire results and key findings on task execution time, respectively.

V. DISCUSSION

This study provides valuable insights into the impact of various control modalities on HRC, focusing on trust, perceived intelligence, safety, cognitive workload, and task execution time for operators with limited robotics expertise. The results highlight the superiority of hand gestures and voice commands over smartPAD controls, emphasizing the benefits of natural and intuitive interfaces. These findings are particularly relevant for micromanagement tasks, such as maintenance operations, where these interaction methods significantly improve performance and user experience.

Key outcomes include a significantly higher *level of trust* when the system is controlled by gestures and voice commands compared to the smartPAD. The trust score increased by 50%, from a mean of 0.60 with the smartPAD to 0.90 with gestures and voice commands. This improvement suggests that users experience greater confidence and security when

[2]Developed by Abele Ingenieure GmbH, a project partner

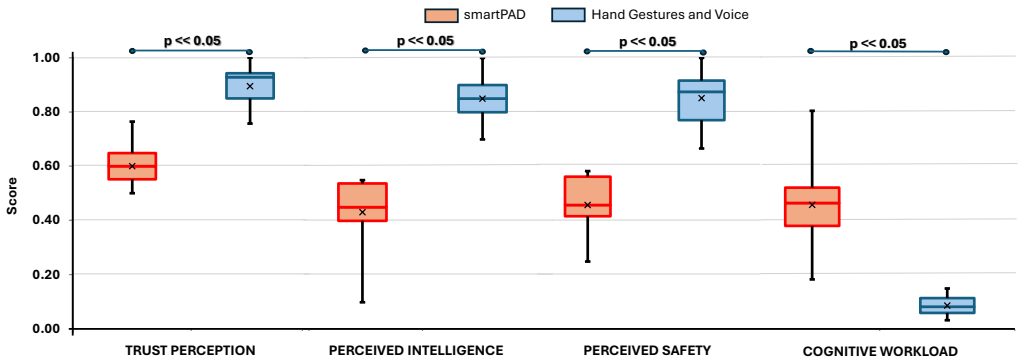

Fig. 7: Comparison of interaction metrics between smartPAD and combined gesture-voice control is shown with normalized scores (0-1) for trust, intelligence, safety, and cognitive workload. All differences are statistically significant ($p \ll 0.05$), with combined gesture-voice control consistently outperforming smartPAD across all metrics.

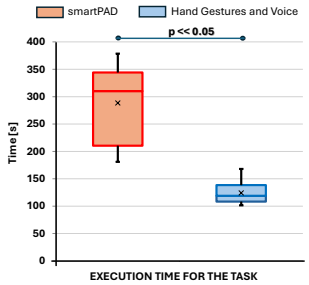

Fig. 8: Task execution time (in seconds) for smartPAD compared to gesture and voice control.

interacting with systems through more natural communication methods, thereby bridging the gap between human operators and machines. Participants also perceived the system as more intelligent and safer with gestures and voice commands. Specifically, the *perceived intelligence* score rose by 97.67%, from a mean of 0.43 with the smartPAD to 0.85 with gestures and voice commands. Similarly, the *perceived safety* score increased by 84.78%, from a mean of 0.46 with the smartPAD to 0.85 with gestures and voice commands. This heightened perception of intelligence likely stems from the intuitive and seamless nature of these interactions, aligning with user expectations for advanced systems, while the increased sense of safety indicates participants felt more in control and reassured by the immediate feedback provided by gestures and voice commands. *Cognitive workload* was significantly higher with the smartPAD, with a mean score of 0.46, compared to a substantially lower mean score of 0.09 for gestures and voice commands, representing an 80.43% reduction in cognitive effort. The higher cognitive workload with the smartPAD reflects the mental strain caused by its less intuitive control interface, which requires translating intended actions into device-mediated commands. In contrast, gestures and voice commands allow for more intuitive communication, reducing cognitive demands and mental strain. Finally, *task execution times* were notably shorter with hand gestures and voice commands compared to the smartPAD. Tasks were completed in an average of 288.45 seconds using the smartPAD, while the mean time for gestures and voice commands was 124.20 seconds, representing a 56.99% reduction in task completion time. This improvement in efficiency is driven by a reduced cognitive workload, increased trust, enhanced perception of system intelligence and safety, as well as optimized path planning and automated execution. These reductions in task execution time can significantly enhance overall performance in manufacturing processes, increasing productivity and streamlining operations.

VI. CONCLUSION

This study introduces a novel framework for robot trajectory control using hand gestures and voice commands, enhancing natural interactions, user confidence, and operational efficiency. Designed to build trust, reduce cognitive workload, and minimize task execution time—especially for operators with limited robotics expertise—the framework ensures seamless integration with ongoing processes, allowing tasks like transporting and draping to resume immediately post-maintenance. Results show that this gesture- and voice-based control significantly improves trust, perceived intelligence, and safety, while reducing cognitive workload and execution times compared to traditional smartPAD controls, thereby enhancing overall manufacturing efficiency.

The framework shows promise but has limitations: it supports only six predefined gestures, is affected by voice command pronunciation issues, and is restricted to single-hand gestures. Future improvements should include enabling more granular, multi-directional movements, adding two-hand gestures for complex tasks, enhancing voice recognition for varied pronunciations, and integrating real-time verbal feedback during robot motion to improve performance.

ACKNOWLEDGMENTS

The authors sincerely thank Dr. Alfons Schuster of the DLR - German Aerospace Center for his support in installing and setting up the industrial camera. They also extend their gratitude to Abele Ingenieure GmbH for their expertise in designing the gripper.

References

[1] L. Wang, R. Gao, J. Váncza, J. Krüger, X. V. Wang, S. Makris, and G. Chryssolouris, "Symbiotic human-robot collaborative assembly," *CIRP annals*, vol. 68, no. 2, pp. 701–726, 2019.

[2] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[3] M. Paliga, "Human–cobot interaction fluency and cobot operators' job performance. the mediating role of work engagement: A survey," *Robotics and Autonomous Systems*, vol. 155, p. 104191, 2022.

[4] K. E. Schaefer, J. Y. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems," *Human factors*, vol. 58, no. 3, pp. 377–400, 2016.

[5] B. M. Muir and N. Moray, "Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[6] J. Sauer, A. Chavaillaz, and D. Wastell, "Experience of automation failures in training: effects on trust, automation bias, complacency and performance," *Ergonomics*, vol. 59, no. 6, pp. 767–780, 2016.

[7] F. Chen, J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, A. Khawaji, D. Conway, F. Chen, J. Zhou, Y. Wang, *et al.*, "Trust and cognitive load," *Robust multimodal cognitive load measurement*, pp. 195–214, 2016.

[8] A. Kalatzis, S. Rahman, V. Girishan Prabhu, L. Stanley, and M. Wittie, "A multimodal approach to investigate the role of cognitive workload and user interfaces in human-robot collaboration," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 5–14.

[9] Y. Li, X. Zhou, X. Jiang, F. Fan, and B. Song, "How service robots' human-like appearance impacts consumer trust: a study across diverse cultures and service settings," *International Journal of Contemporary Hospitality Management*, 2024.

[10] C. Moro, S. Lin, G. Nejat, and A. Mihailidis, "Social robots and seniors: A comparative study on the influence of dynamic social features on human–robot interaction," *International Journal of Social Robotics*, vol. 11, pp. 5–24, 2019.

[11] A. C. Simões, A. Pinto, J. Santos, S. Pinheiro, and D. Romero, "Designing human-robot collaboration (hrc) workspaces in industrial settings: A systematic literature review," *Journal of Manufacturing Systems*, vol. 62, pp. 28–43, 2022.

[12] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 141–148.

[13] R. Silva, M. Faria, F. S. Melo, and M. Veloso, "Adaptive indirect control through communication in collaborative human-robot interaction," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3617–3622.

[14] L. Lu, Z. Xie, H. Wang, B. Su, and X. X. Edward P. Fitts, "Assessing workers' mental stress in hand-over activities during human-robot collaboration," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 66, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2022, pp. 537–541.

[15] G. Campagna, D. Chrysostomou, and M. Rehm, "Analysis of facial features for trust evaluation in industrial human-robot collaboration," in *2024 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE, 2024, pp. 1–6.

[16] G. Campagna, M. Dadgostar, D. Chrysostomou, and M. Rehm, "A data-driven approach utilizing body motion data for trust evaluation in industrial human-robot collaboration," in *33rd IEEE International Conference on Robot and Human Interactive Communication, IEEE RO-MAN 2024*. IEEE, 2024.

[17] P. Neto, M. Simão, N. Mendes, and M. Safeea, "Gesture-based human-robot interaction for human assistance in manufacturing," *The International Journal of Advanced Manufacturing Technology*, vol. 101, pp. 119–135, 2019.

[18] C. Li, X. Zhang, D. Chrysostomou, and H. Yang, "Tod4ir: A humanised task-oriented dialogue system for industrial robots," *IEEE Access*, vol. 10, pp. 91 631–91 649, 2022.

[19] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the "trust perception scale-hri"," in *Robust intelligence and trust in autonomous systems*. Springer, 2016, pp. 191–218.

[20] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[21] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, pp. 71–81, 2009.

[22] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[23] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1145–1153.

[24] G. Sung, K. Sokal, E. Uboweja, V. Bazarevsky, J. Baccash, E. G. Bazavan, C.-L. Chang, and M. Grundmann, "On-device real-time hand gesture recognition," *arXiv preprint arXiv:2111.00038*, 2021.

[25] T. Haase and M. Schönheits, "Towards context-aware natural language understanding in human-robot-collaboration," in *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2021, pp. 1648–1653.

[26] A. Gottardi, E. Pagello, N. Castaman, E. Menegatti, *et al.*, "Human-robot task and motion planning in an industrial application," in *IEEE/RSJ IROS Workshop on Task and Motion Planning: from Theory to Practice*, 2023.

[27] J. J. Kuffner and S. M. LaValle, "Rrt-connect: An efficient approach to single-query path planning," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, vol. 2. IEEE, 2000, pp. 995–1001.

[28] N. Miotto, A. Gottardi, N. Castaman, and E. Menegatti, "Collision-free volume estimation algorithm for robot motion deformation," in *2023 21st International Conference on Advanced Robotics (ICAR)*. IEEE, 2023, pp. 348–354.

[29] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, "I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system," *Human factors*, vol. 55, no. 3, pp. 520–534, 2013.